

# Between-Region Genetic Divergence Reflects the Mode and Tempo of Tumor Evolution

Ruping Sun<sup>1,2,3§</sup>, Zheng Hu<sup>1,2,3§</sup>, Andrea Sottoriva<sup>4</sup>, Trevor A. Graham<sup>5</sup>, Arbel Harpak<sup>6</sup>, Zhicheng Ma<sup>1,2,3</sup>, Jared M. Fischer<sup>7</sup>, Darryl Shibata<sup>8</sup>, Christina Curtis<sup>1,2,3\*</sup>

## Affiliations

<sup>1</sup> Department of Medicine Stanford University School of Medicine, Stanford, CA 94305

<sup>2</sup> Department of Genetics Stanford University School of Medicine, Stanford, CA 94305

<sup>3</sup> Stanford Cancer Institute, Stanford University School of Medicine, Stanford, CA 94305

<sup>4</sup> Centre for Evolution and Cancer, The Institute of Cancer Research, London, SM2 5NG, UK

<sup>5</sup> Barts Cancer Institute, Queen Mary University of London, London, John Vane Science Centre, Charterhouse Square, London EC1M 6BQ, UK

<sup>6</sup> Department of Biology, Stanford University, Stanford, CA 94305

<sup>7</sup> Oregon Health and Science University, Department of Molecular and Medical Genetics, 3181 SW Sam Jackson Park Road, Portland, OR 97239

<sup>8</sup> Department of Pathology, Keck School of Medicine of the University of Southern California, Los Angeles, CA 90033

§ These authors contributed equally to this work

\* Correspondence should be addressed to: Christina Curtis, Stanford University School of Medicine, 265 Campus Drive, Lorry Lokey Building Suite G2120C, Stanford, CA 94305 Tel: 650-498-9943, Email: [cncurtis@stanford.edu](mailto:cncurtis@stanford.edu)

## **Abstract**

Given the implications of tumor dynamics for precision medicine, there is a need to systematically characterize the mode of evolution across diverse solid tumor types. In particular, methods to infer the role of natural selection within established human tumors are lacking. By simulating spatial tumor growth under different evolutionary modes and examining patterns of between-region subclonal genetic divergence from multi-region sequencing (MRS) data, we demonstrate that it is feasible to distinguish tumors driven by strong positive subclonal selection from those evolving neutrally or under weak selection, as the latter fail to dramatically alter subclonal composition. We developed a classifier based on measures of between-region subclonal genetic divergence and projected patient data into model space, revealing different modes of evolution both within and between solid tumor types. Our findings have broad implications for how human tumors progress, accumulate intra-tumor heterogeneity, and ultimately how they may be more effectively treated.

# Introduction

The multistage model of carcinogenesis described in the early 1950s<sup>1,2</sup> and Nowell's 1976 perspective piece on the clonal evolution of tumor cells<sup>3</sup> provided a conceptual framework for understanding tumor progression. These and other studies<sup>4,5</sup> were foundational in defining the elements of somatic evolution. However, the evolutionary dynamics that govern tumor initiation and subsequent growth after transformation remain poorly understood. Moreover, the distinction between stages is often blurred since tumorigenesis is largely occult often taking place over decades<sup>6,7</sup> where lesions are only detected once they achieve a certain size or cause symptoms.

Evolution is the product of three major underlying processes: mutation, selection and genetic drift<sup>8</sup>. Mutations are readily measured in human tumors, and it is generally assumed that ongoing strong selection governs the growth of an established tumor after transformation, leaving a detectable signal on the genome, where the acquisition of additional 'drivers' results in multiple selective sweeps<sup>9,10</sup>. In this scenario, driver mutations accompanied by numerous hitchhiking passengers can attain high frequency and manifest as 'subclonal clusters' in bulk tumor sequencing data<sup>10</sup>. This led to the development of a suite of methods aimed at inferring subclonal clusters. However, inference of the number of subclones and their proportions from bulk tumor sequencing is a non-trivial task with the solution non-identifiable under most conditions<sup>11-14</sup>. Drift can also cause extensive intra-tumor heterogeneity (ITH) that may be difficult to distinguish from selection without appropriate population genetics methods. For example, we proposed and tested several predictions of a Big Bang model of colorectal tumor growth, wherein *after* transformation, the tumor grows as a single terminal expansion populated by a large number of heterogeneous—and effectively equally fit—subclones<sup>15</sup>. In this model, most detectable subclonal (private) alterations arise early during growth. While post-transformation selection could be detected in these colorectal tumors, it was often too weak to alter tumor subclonal architecture. Rather, patterns of ITH were suggestive of effectively-neutral evolution.

Other studies have since corroborated 'Big Bang' dynamics in colorectal tumors<sup>16-19</sup>. Additionally, neutral evolution was reported in hepatocellular carcinoma via in depth multi-region profiling<sup>20</sup>. Williams *et al.* further investigated evidence for neutral evolution in multiple solid tumors using bulk single sample sequencing data compared to a theoretical null neutral model<sup>21</sup>. However, as we show, this task is better powered using MRS, which captures additional features of genetic diversity.

Progression modes and tempos differ between neutrally evolving tumors and those tumors with post-transformation selection. Hence, there remains a need for the systematic evaluation of different modes of evolution in diverse solid tumors within a population genetics framework. As selection is complex, it is instructive to initially focus on the commonly assumed scenario of strong positive selection after transformation and contrast this with a neutral model. We leverage the fact that spatiotemporal patterns of genetic variation among cancer cell populations and in particular their variant allele frequency (VAF) distributions (also known as the site frequency spectrum or SFS)<sup>22</sup> derived from next generation sequencing (NGS) can be used to test hypotheses about the underlying evolutionary processes, including the strength of selection and extent of genetic drift. To this end, we simulated spatial tumor growth under different modes of evolution and trained a classifier based on ITH metrics derived from the SFS to discriminate between these scenarios. By projecting MRS data from various solid tumors into model space, we categorize their patient-specific evolutionary dynamics.

## Results

### Spatial simulation of distinct modes of tumor evolution

To investigate how different modes of tumor evolution influence the SFS from bulk sequencing data, as well as the power to detect signals of positive selection, we developed an agent-based model of spatial tumor growth (parameters reported in **Supplementary Table 1**). Within this framework, we simulated various modes of tumor evolution, including a neutral model and an alternate neutral model based on cancer stem cell (CSC) driven growth (neutral-CSC). We also simulated various levels of positive selection ( $s=0.01, 0.02, 0.03, 0.05, 0.1$ ), such that the acquisition of advantageous mutations alters the cell birth-death rate according to the selection coefficient,  $s$  (**Figure 1, Supplementary Figure 1, Methods**). In all models random neutral point mutations arise via a Poisson process during each cell division. Virtual tumor growth is simulated via the expansion of deme<sup>23</sup> subpopulations (i.e. neighborhoods of 5-10k cells) within a defined 3D lattice, and cells within each deme are well-mixed and replicate via a random branching process. By recording mutational lineages as the tumor expands and subsequently virtually sampling the 'final' tumor as is done experimentally after resection, we evaluate differences in the SFS arising under different levels of selection, and the utility of different tissue sampling strategies (**Figure 1a**). Thus, we model spatial tumor growth and the inherent stochasticity of this process while accounting for the truncated SFS derived from bulk sequencing due to the large number of rare subclones that are not sampled or below detection limits. This facilitates comparisons with data derived from patient tumors analyzed within a sensitive pipeline for calling somatic single nucleotide variants (SSNVs) from MRS (**Figure 1b, Supplementary Figure 2, Methods**). A summary of terminology is provided in **Supplementary Table 2**.

Spatial subclone composition and the distribution of subclonal VAFs derived from MRS ( $n=2, 4$  and  $8$  regions) of 'virtual' tumors differed dramatically depending on the mode of evolution, as illustrated for representative virtual tumors (**Figure 2ab, Supplementary Figure 3**). In particular, under stronger selection ( $s \geq 0.02$ ), multiple subclone expansions occur in different regions of the virtual tumor, as shown in the clone map (**Figure 2a**). Likewise, multiple peaks (mutational clusters) were observed in the SFS histograms due to the enrichment of high frequency ( $\text{VAF} > 0.2$ ) subclonal SSNVs under stronger selection (**Figure 2b** and shown schematically in **Supplementary Figure 4**), which were largely region-specific, reflecting elevated genetic divergence. Indeed, subclonal selection typically resulted in detectable differences in the SFS histograms from different tumor regions. In contrast, under neutral growth, a neutral CSC-like model where only a subset of cells have unlimited proliferative potential (equivalent to a smaller deme size), or weak selection ( $s=0.01$ ), subclonal composition is preserved in the final tumor. The SFS for these three modes were generally similar between regions consisting of two 'mutational clusters', namely a public cluster centered at  $0.5$  VAF composed of mutations that occurred prior to transformation and present in all tumor cells (fixed) and a right skewed distribution of private (subclonal) mutations at low VAF ( $< 0.25$ ) (**Figure 2b**), where their detection depends on sequencing depth. Importantly, MRS but not single-sample sequencing enables the identification of private SSNVs present at high frequency in one or a few regions, but subclonal in the entire tumor (**Supplementary Figure 4**). Indeed, at least two spatially separated regions are needed to accurately distinguish public SSNVs in solid tumors, as mutations that are subclonal in the whole tumor can appear 'clonal' within some samples due to sampling bias<sup>24</sup>. In each of the modes, over 70% of subclonal SSNVs were region-specific due to spatial constraints during virtual tumor

expansion. However, selection increased the fraction of high frequency ( $\text{VAF} > 0.2$ ) region-specific subclonal SSNVs out of all region-specific subclonal SSNVs ( $\text{VAF} > 0.08$ ) (**fHrs**) (**Supplementary Table 4**). Hence, MRS aids the identification of subclonal SSNVs that reflect the dynamics of clonal expansion after tumor transformation, whereas clonal SSNVs are not informative in this regard.

To quantify the extent of ITH defined as between-region genetic divergence based on *subclonal* SSNVs (identified through MRS) under different levels of selection, we employed the following metrics (Methods) in addition to **fHrs** (defined above):

**fHsub** – fraction of subclonal SSNVs ( $\text{VAF} > 0.08$ ) with high frequency ( $\text{VAF} > 0.2$ ).

**F<sub>ST</sub>** (Fixation index) – a measure of genetic divergence between regions<sup>25</sup>.

**KSD** (Kolmogorov-Smirnov distance) – dissimilarity of the SFS between regions.

As expected, fHrs and fHsub were correlated, as were other features, albeit to a lesser extent (**Supplementary Figure 5**). All of the statistics increased in value under stronger selection ( $s \geq 0.02$ ) relative to the neutral/neutral-CSC/weak selection ( $s = 0.01$ ) models. This suggests that selection causes characteristic and detectable genetic divergence between regions when it fails to result in complete sweeps (**Figure 2b**, **Supplementary Table 4**).

We further explored the relationship between different modes of evolution and genetic divergence captured by MRS ( $n = 2, 4, 8$  regions) and single sample sequencing at various depths (80-640x) (**Figure 2c**, **Supplementary Figures 6-7**, Methods). For reference, the theoretical cumulative SFS assuming neutral exponential growth in a well-mixed population<sup>21,26</sup> (referred to as the theoretical neutral SFS) is also shown. Differences in the SFS were evident such that tumors simulated under higher selection ( $s \geq 0.02$ ) typically fell above the theoretical neutral SFS, whereas the remaining modes generally traced or fell below this curve. The variability in the SFS within individual modes highlights the importance of stochastic simulations.

To compare the utility of single sample data versus MRS, we computed the ratio of the area under the cumulative SFS (based on the pooled VAF for MRS) to the area under the theoretical neutral SFS (**rAUC**) as this is applicable to both single sample and MRS. Comparison of the rAUC for virtual tumors simulated under different modes demonstrates the challenge of distinguishing between  $s > 0.05$  or  $s \leq 0.01$  (including the neutral and neutral-CSC models) with a single sample, even at high depth, whereas better separation is achieved with even one additional region (**Supplementary Figure 8**). This is also reflected in comparisons of the sensitivity and specificity to distinguish alternative models from the simulated neutral model based on the rAUC (**Supplementary Figure 9a**). Whereas power increased with selection intensity ( $s = 0.05$ - $0.1$ ) and the number of regions ( $n = 2$ - $8$ ), this was not the case for increased depth alone due to sampling bias and the inability to capture regionally localized high frequency subclonal mutations that arose under strong selection (**Supplementary Figure 9b**, Methods). In contrast, metrics that capture between-region ITH such as fHsub are better able to distinguish a specific alternate model than rAUC. Of note,  $s = 0.01$  could not be distinguished from the simulated neutral model. The neutral-CSC model is also similar to the ‘vanilla’ neutral model, but generates localized diversity. Thus, we refer to these three modes as effectively-neutral, since the population dynamics of such nearly neutral mutations are virtually equivalent to those of neutral mutations<sup>27,28</sup>. Similarly, it was not feasible to distinguish the SFS under different levels of elevated selection ( $s \geq 0.02$ ) (**Supplementary Figure 5**). Many factors can dampen signals of selection as in the case of strong, but less frequent ‘drivers’ that are very rare or occur late without sufficient time to expand (**Supplementary Figure 10**). As such, we focus on effective

neutrality and strong selection ( $s \geq 0.02$ ), but present results from all modes for completeness.

### **The site frequency spectrum reflects tumor growth dynamics**

In order to evaluate the SFS in patient samples, we first analyzed MRS data from colorectal adenocarcinomas sampled from two regions (COAD, taken >3 cm apart)<sup>15</sup> with high purity (72-96%) and adequate coverage (80-120X median WES depth) (**Supplementary Figures 11-12**). We devised a MuTect-based Variant Assurance Pipeline (VAP) to enable the sensitive and accurate detection of subclonal SSNVs from MRS (**Supplementary Figures 2, 13, Methods and Supplementary Note**). The observed VAF estimates were adjusted for sample purity and local copy number, enabling pairwise comparisons between tumor regions, and throughout we refer to adjusted VAFs as VAFs (**Supplementary Figures 14-15**). As noted above, the SFS histograms appear bimodal for both regions, as shown for representative tumors from the major COAD subgroups, namely microsatellite instable (MSI-H), microsatellite stable/chromosomally stable MSS/CIN- and MSS/CIN+<sup>29</sup> (**Figure 3a**). A peak centered at a VAF of 0.5 was observed in all tumors with constituent mutations that were present at similar frequencies in the left and right samples (**Figure 3b**). This VAF cluster primarily represents *public* mutations present in the founding tumor cell. Whereas private high-VAF (0.2-0.4) SSNVs were infrequent, low frequency subclonal SSNVs (VAF<0.2) were common and generally region-specific despite having similar VAF, suggesting that mutation frequency is not a reliable surrogate for subclone identity. Similar patterns were observed in additional cancers and an adenoma (**Supplementary Figure 15**). We computed the five ITH metrics, which exhibited low or intermediate values for COADs M, O, and U comparable to those noted in ‘virtual’ tumors under effectively-neutral growth. In contrast, tumors G, N, W, and adenoma S exhibited higher values, similar to those noted in ‘virtual’ tumors subject to selection (**Figure 3, Supplementary Tables 4-5**).

We further evaluated the genetic divergence within a clonal *in vivo* tumor growth model by generating single cell expansions from mismatch repair (MMR) deficient COAD cell lines followed by xenotransplantation into opposite flanks of immune compromised mice and WES of the resultant tumors (Methods). In both technical replicates and independent cell line experiments, the data yielded SFS histograms that lacked enrichment for high-frequency private SSNVs (**Supplementary Figures 16-17**). Additionally, the corresponding ITH metrics were congruent with effectively-neutral growth, as might be expected for fully transformed cells that do not require further alterations to propagate tumor growth.

### **VAF clusters do not necessarily capture subclone identity**

Existing computational methods to infer tumor subclonal architecture from bulk sequencing data exploit the observation that SSNVs cluster around several distinct VAF modes or ‘clonal clusters’<sup>10-13,30</sup>. These methods aim to assign ‘subclone’ identity based on the assumption that mutations with similar frequencies are in the same cell and that a limited number of dominant subclones underwent clonal expansion<sup>9,11,31</sup>. However, mutational clusters do not guarantee unique lineages, and therefore do not necessarily capture clonal identity. In addition, subclone architecture is influenced by the mode of tumor evolution and spatial constraints. Indeed, visual inspection of the SFS histograms and scatterplots from the bi-sampled COAD dataset revealed that in all cases, the majority of subclonal SSNVs with VAF<0.2 were region-specific (**Figures 3, Supplementary Figure 15**). This suggests that mutations grouped based on their VAF do not correspond to unique clones. To evaluate subclonal architecture at higher

resolution, we performed WES on five individual COAD glands and bulk samples from two distant tumor regions of a representative cancer (COAD O). The private mutations specific to either bulk sample (OA or OB, **Figure 4a, b**) were only detected in glands from the same tumor region ( $p=1E-10$ , Fisher's exact test) and similar patterns were noted based on targeted sequencing of private SSNVs in multiple individual glands for each of the bi-sampled COADs (**Supplementary Figure 18**). In a subset of single glands from two spatially separated regions, the same SSNVs were detected despite being subclonal in the bulk tumor (**Figure 4b**, green dots), potentially reflecting early subclone mixing<sup>15,19</sup> or sampling of a clone boundary. In contrast, later arising SSNVs were generally restricted to one region, consistent with spatial constraints during expansion. SSNVs specific to bulk sample OA (VAF < 0.2) were detected in different combinations of single glands with VAF > 0.2, suggesting that distinct lineages can have similar VAFs in the bulk tumor. Reconstruction of a possible phylogenetic tree using LICHeE<sup>32</sup> also revealed subclone spatial segregation, where essentially every gland within a bulk region is a subclone (**Figure 4d**), emphasizing the star-like phylogeny predicted for a neutrally growing population<sup>26</sup> (**Supplementary Figure 19**). WES of single glands from COAD U yielded similar results (**Supplementary Figure 20**).

We further reasoned that a 'true' clone should form a cluster that persists (e.g. mutations remain grouped), irrespective of the inclusion of data from additional regions. We evaluated this in other solid tumors by analyzing published MRS datasets for esophageal carcinoma (ESCA)<sup>33</sup>, lung adenocarcinoma (LUAD)<sup>34</sup>, non-small cell lung cancer (NSCLC)<sup>35</sup>, glioma (GLM)<sup>36</sup> and glioblastoma (GBM)<sup>37</sup> (**Supplementary Figure 2, 11, Supplementary Table 3, Methods**). Application of SciClone<sup>13</sup> to MRS data from several representative tumors (COAD-O, ESCA-8, LUAD-4990 for which 2, 3 and 4 regions were available, respectively) consistently resulted in the separation of subclonal clusters when data from additional regions were included in the analysis (**Figure 4c, Supplementary Figures 21-23**). Whereas SSNVs in the subclonal clusters did not remain grouped, those in the clonal clusters did ( $p=0.0003$ , Fisher's exact test), consistent with them being in the founding clone. A persistent mutational cluster in LUAD-4990 was detected through the analysis of 4 regions, potentially corresponding to a subclone that arose under selection (**Supplementary Figure 22**). Collectively, these results illustrate conceptual challenges in inferring subclonal architecture from bulk sequencing VAF data alone.

## Distinguishing the mode and tempo of solid tumor evolution

We next evaluated genetic divergence based on MRS of treatment naïve primary tumors, including COAD, ESCA, LUAD, LUSC, and GBM relative to those observed in virtual tumors under different modes. Non-hypermutated GBMs ( $n=2$ ) and gliomas ( $n=2$ ) obtained pre- and post-treatment with temazolamide, a mutagenic alkylating agent assumed to impose a positive selective pressure<sup>36</sup>, were included as positive controls. Additionally, matched Barrett's esophageal (BE) lesions and adenocarcinomas from two patients (BE-ESCA-4 and BE-ESCA-14) were included as positive controls, since selection is expected during progression from a pre-malignant lesion. The degree of deviation of the pooled cumulative SFS above the theoretical neutral curve highlights differences in selection across tumor types (**Figure 5a**). As predicted, each of the positive controls exhibited cumulative SFSs above the neutral curve, consistent with strong selection. In contrast, deviation below the theoretical neutral curve is indicative of spatial constraints, as illustrated by simulating smaller deme sizes (0.5-1k vs. 5-10k), where the ability to distinguish selection from effective neutrality was reduced (**Supplementary Figures 24-25**). Such strong spatial constraints result in infrequent sharing of subclonal mutations between regions (fShr, **Supplementary Table 4-5**), a

pattern inconsistent with most patient tumors ( $p < 2.2\text{e-}16$ , Wilcoxon rank sum test), suggesting that larger deme size better reflects the patient data.

COAD-M and ESCA-14 exhibited bimodal SFS histograms with scant enrichment for high frequency private SSNVs, most consistent with patterns of effective neutrality (**Figure 5b**). In contrast, COAD-N and LUAD-270 exhibited modest enrichment for such SSNVs, whereas this was more striking in ESCA-8 and LUAD-4990. Despite the lower number of SSNVs in treatment-naïve primary GBMs, enrichment of high frequency private SSNVs was evident and similar to that noted in the primary versus post-treatment recurrence (**Figure 5b**).

The five ITH metrics were calculated for primary solid tumors, paired pre- and post-temazolamide treated gliomas and GBMs (positive controls) and BE-ESCA pairs (positive controls), as well as virtual tumors simulated under various evolutionary modes (**Figure 6a**). Amongst the virtual tumors, all five metrics increased markedly under selection ( $s \geq 0.02$ ) relative to effective neutrality. The primary COADs and ESCAs tended to exhibit lower detectable divergence than lung and brain cancers, which were lower than the temazolamide treated positive controls.

The SFS is commonly used in population genetics<sup>22,38</sup> and it is appreciated that tests of neutrality based on a single summary statistic can be difficult to establish, whereas composite metrics can aid the detection of selection<sup>39</sup>. Given the multi-faceted nature of ITH and the noise in real data, we reasoned that the major components of the ITH metrics would capture complementary aspects of subclonal genetic divergence. Independent component analysis (ICA) using the five ITH metrics revealed two distinct clusters, corresponding to selection with  $s \geq 0.02$  and neutral/weak selection ( $s = 0.01$ )/neutral-CSC (**Figure 6b**). A support vector machine (SVM) was trained on the two independent components (ICs) to discriminate between effectively-neutral evolution (3 modes with  $s \leq 0.01$ ) and selection (4 modes with  $s \geq 0.02$ ). The SVM based on the ICs performed better than individual ITH metrics and although models using two or more ITH metrics performed well (**Supplementary Figures 26-27**), we adopted the two ICs to survey genetic divergence in patient samples.

We then classified patient tumors and visualized them in model space (**Figure 6b**, **Supplementary Figures 28-29**, **Supplementary Table 5**), revealing trends with respect to the mode of evolution in a given tumor type, despite patient to patient variability. For example, COADs exhibited both effective neutrality as well as selection, as did ESCAs. In contrast, lung and brain tumors tended to show stronger signals of selection. In total, 5 primary tumors were categorized as being compatible with effective-neutrality and 12 with selection, whereas only 3 did not robustly fit either scenario. As expected, all four pre- versus post-temazolamide treated GBMs and gliomas were most compatible with strong positive selection and several appear as outliers on the ICA, potentially because the full impact of treatment is not modeled (**Figure 6b**). The paired BE-ESCA cases (ESCA\_BE-14 and ESCA\_BE-4) exhibited patterns consistent with selection during tumorigenesis, followed by effectively-neutral growth of the primary (ESCA-14 and ESCA-4). Patterns of genetic divergence in multiple BE lesions from patient 4 (BE-4) were similarly indicative of selection (**Supplementary Figure 30**). Importantly, irrespective of whether WGS or WES data was used, the classification was the same indicating that WES is adequate for this task given sufficient subclonal SSNVs (**Figure 6b**).

Positive selection for ‘drivers’ during tumor expansion is expected to be associated with an increase in the rate of private SSNVs at more functional (MF) relative to less functional (LF) sites<sup>40</sup>. Amongst primary tumors, the dMF/dLF ratio was positively



correlated with several ITH metrics, e.g., fHsub, FST, and rAUC (**Figure 6c**). This suggests a general trend between selection and the levels of detectable between-region genetic divergence, although specific patterns could be model dependent (**Supplementary Figure 31**). Conversely, the fold enrichment for driver genes amongst non-silent public SSNVs was negatively correlated with fHsub, consistent with a greater number of public drivers in tumors characterized by effectively-neutral growth (**Supplementary Figure 32**). Hence, these results corroborate our finding that patterns of genetic divergence in MRS inform the mode and drivers of tumor growth.

## Discussion

Here we show that tumors evolving near neutrally or through strong selection exhibit fundamentally different patterns of ITH and that these can be distinguished via MRS. Further, we developed a classification framework based on features of the SFS that capture between-region subclonal divergence and applied this to publicly available MRS data, revealing different modes of evolution within and between solid tumor types. We note that compatibility with effective neutrality does not necessarily imply the complete absence of selection. Rather, selection may have been weak or variable throughout tumor growth, but the overall patterns do not deviate significantly from those expected under a neutral model. The timing of a mutation is also critical since within a rapidly expanding adaptive population, only mutations that occur early are likely to be ‘fixed’ in relevant time frames and detectable by NGS, even if they are under strong positive selection, whereas partial sweeps are potentially common<sup>41</sup>. The lack of evidence for ongoing stringent selection in some of the tumors examined here is congruent with a Big Bang model of effectively-neutral tumor growth where the tumor grows as a single expansion with selection uniformly conferred by common drivers in the first tumor cell<sup>15</sup>.

The finding that human tumors can be categorized into different modes of evolution has implications for defining the ‘drivers’ of growth and treatment strategies. For example, near-neutrally evolving tumors show enrichment for drivers amongst public SSNVs, and it is potentially most efficacious to target these truncal mutations. While most *detectable* ITH occurs early during effectively-neutral growth, the large number of heterogeneous subclones that fall below detection limits increases the chance that pre-existing treatment resistant variants are present. In contrast, putatively functional private variants were enriched amongst tumors characterized by ongoing positive selection, suggesting these may represent relevant targets.

Our findings also inform practical guidelines for studies of tumor evolution. For example, we show that while at least two regions are required to robustly distinguish public versus private alterations, inclusion of sequencing data from additional regions yielded greater discrimination between different modes of evolution and was more informative than deeper sequencing of a single sample. Even under strong spatial constraints such as small (0.5-1k) deme size, where the efficacy of selection is impeded, sequencing additional regions should aid the detection of selection. Improved sensitivity to distinguish different modes of evolution may be achieved by modeling the distinct architecture and microenvironments in different tissues, although these are as of yet poorly understood<sup>42</sup>. It will also be important to understand the contribution of deleterious passenger alterations<sup>43</sup> and clonal cooperation<sup>44,45</sup> to tumor dynamics, as well as to evaluate more complex modes of selection in human tumors. Thus, although MRS does not fully resolve the SFS, it nonetheless captures global and local genetic divergence, enabling the detection of signals of selection in individual under certain conditions.

## Online Methods

### Multi-region sequencing studies

We evaluated patterns of ITH in several publicly available MRS datasets spanning multiple tumor types, including colorectal adenoma/adenocarcinoma (1 adenoma, 6 COAD patients)<sup>15</sup>, esophageal carcinoma/Barrett's esophageal (ESCA, 3 patients)<sup>33</sup>, lung adenocarcinoma (LUAD, 4 patients)<sup>34</sup>, non-small cell lung cancer (NSCLC, 1 patient)<sup>35</sup>, glioma (GLM, 3 patients)<sup>36</sup>, and glioblastoma (GBM, 2 patients)<sup>37</sup>, numbers refer to cases with MRS data that passed QC. The study accession IDs and list of samples that met coverage and purity requirements are reported in **Supplementary Table 3**. Details on sequencing depth and purity are provided in **Supplementary Figure 11**. All samples were analysed using a custom pipeline (**Supplementary Note**) to enable the sensitive detection of private SSNVs and standardized comparisons across cohorts, as detailed below.

### Single gland whole-exome sequencing

Building on our prior description of multi-region WES of colorectal tumors and targeted single gland sequencing, we performed WES of multiple single glands from two tumors in this study (**Figures 4, Supplementary Figure 20**) on the Illumina platform using the Agilent SureSelect 2.0 or Illumina NRCE kit. Samples were collected under an IRB approved protocol as previously described<sup>15</sup>. The single gland WES data were analyzed using the same pipeline as was applied to bulk tumor regions. Intersection plots for SSNVs found in bulk regions and single-glands were generated based on mutations that were i) covered by at least 20 reads in each sample; ii) with a VAF above 1.5% in the bulk sample or above 15% in the single-glands; and iii) do not derive from regions with varying patterns of LOH amongst samples.

### *In vivo* modeling of colorectal tumor growth

Cells were expanded *in vitro* and a single 'founding' cell from this population was cloned and expanded to ~6 million (M) cells prior to transplantation of ~1M cells into the right and left flanks of a NSG mouse (HCT116) or a Nude (Nu/Nu) mouse (LoVo), where tumors were allowed to develop to a size of ~1 billion cells (1 cm<sup>3</sup>) before being sampled and subject to WES (**Figures 3, Supplementary Figure 16-17**). The HCT116 and LoVo MMR-deficient COAD cell lines were obtained from the ATCC (authenticated using cytochrome C oxidase I assays and STR typing and tested for mycoplasma contamination) and cultured under standard conditions. Tissue was collected separately from the right and left tumors and DNA was extracted for WES using the Illumina TruSeq Exome kit, as was DNA from the first passage population (a polyclonal tissue culture for HCT116 and a polyclonal xenograft sample for LoVo), which were employed as a reference for detecting SSNVs and for CNA estimation.

### Somatic SNV calling, SCNA detection and VAF adjustment

To facilitate quantitative comparisons of the SFS, we devised a unified variant assurance (filtering and rescuing) pipeline (VAP) to achieve balance in sensitivity and specificity when MRS is available such that information can be borrowed across tumor regions. For each raw SNV call by MuTect (v1.1.4, unfiltered)<sup>46</sup>, the read alignment features from all samples was re-inspected in an automated fashion to assess the confidence (in detected samples) and evidence (in un-detected samples) for the alternative allele (**Supplementary Figure 13**). Somatic copy number alterations and tumor purity (*p*) were estimated with *TitanCNA*<sup>47</sup> (version 1.8.0) in exome-seq mode (except for the ESCA dataset where WGS was available). The observed VAF for each

detected somatic SNV was adjusted based on CCF (Cancer Cell Fraction) calculation by taking into account tumor purity, local copy numbers as well as the inferred time ordering between SCNA and SSNV as previously described<sup>31</sup> (**Supplementary Figure 14-15**), in order to enable comparisons of genetic divergence between regions. Additional details for this section, including benchmarking of VAP (**Supplementary Figure 33-35**), can be found in the **Supplementary Note**.

### **Spatial computational modeling of tumor growth dynamics Ino**

We extended our previously described spatial agent-based model<sup>15</sup> to simulate tumor growth and mutation accumulation under different scenarios ranging from neutral evolution to strong selection and compare the SFS of SSNVs arising from 1, 2, 4 and 8 regions sampled from spatially separated quadrants of individual virtual tumors. In this agent-based model, spatial tumor growth is simulated via the expansion of deme subpopulations (composed of 5-10k cells), which mimics the glandular structures often found in epithelial tumors (**Supplementary Table 1**). The deme model is well established for modeling spatially expanding populations<sup>23</sup>. Here, deme subpopulations expand within a defined 3D cubic lattice (Moore neighborhood, 26 neighbors), where demes expand by particular rules of spatial constraints (peripheral growth<sup>48</sup> or alternatively shifting growth<sup>15</sup>) while cells within each deme are well-mixed and grow via a random branching (birth-death) process. The panmixia of cells in the formation of the first deme from a single transformed cell allows for subclone mixing amongst early-arising mutations<sup>15,19</sup>, which can subsequently spread during tumor expansion. Random neutral mutations arise via a Poisson process at each cell division, assuming an infinite sites model.

More specifically, at each time step, we simulate deme division by selecting a deme at random and choosing a neighboring lattice site where the new deme will be placed. We employ a peripheral growth model<sup>48</sup> (**Supplementary Table 1**), where only demes on the surface of the tumor can grow and divide such that a random empty neighbor site was chosen for each newly generated deme. The peripheral growth model is supported by recent studies indicating that cancer cells at the periphery of the tumor exhibit higher proliferative activity than those at the core<sup>42</sup>. We assume a maximum deme size of 10,000 cells in order to minimize the effect of deme structure, which hinders selection. While we focus on this conservative scenario, we also explored the impact of a smaller deme sizes (down to 1,000 cells) (**Supplementary Figures 24-25**). Within the model there is no spatial partition for tumor cells within demes which proliferate via a discrete stochastic birth-and-death process (division rate  $p$  and death rate  $q=1-p$ , the death/birth ratio  $h=q/p$ ), where the first deme is generated by the same process beginning with a single transformed tumor cell. Simple birth-death processes give rise to exponential growth of each deme on average where the growth rate is  $r=\ln(2p)$ . Here we employ the following parameters:  $p=0.55$ ,  $q=0.45$  and thus  $r=\ln(2\times 0.55)\approx 0.1$  as the growth rate of deme expansion, where  $p$  and  $q$  were empirically chosen by assuming a relatively high death versus birth rate ( $h=q/p=0.82$ ) in each cell generation in line with previous estimates in a rapidly growing colorectal cancer metastasis ( $h=0.72$ )<sup>49</sup> and in early tumors ( $h=0.99$ )<sup>50</sup>. Once the deme exceeds the maximum size, the deme will split into two offspring demes via sampling from a binomial distribution  $[N_c, p=0.5]$  where  $N_c$  is the current deme size. During each cell division, the number of neutral passenger mutations that arise in the coding portion of the genome follows a Poisson distribution with mean,  $u$ , where an infinite sites model and constant mutation rate was assumed. Under the null model, all somatic mutations are assumed to be neutral and do not confer a fitness advantage, whereas in the selection models, beneficial mutations (or advantageous mutations) occur stochastically via a Poisson process with mean  $u_b$

during each cell division. Thus, we consider the null neutral model ( $s=0$ ), as well as varying degrees of selection:  $s=0.01, 0.02, 0.03, 0.05$  and  $0.1$ , where  $s$  is the selection coefficient defined by the increase in the cell division rate when a beneficial mutation occurs in the neutral cell lineage. The cell division rate and death rate of a selectively beneficial clone is  $p_b=p \times (1+s)$  and  $q_b=1-p_b=1-p \times (1+s)$ , respectively. The growth rate of a selective lineage within a deme is  $r_b=\ln(2 \times p_b)$ . The parameters employed are reported in **Supplementary Table S1** and include  $u=1.2$  within the 60 Mb of coding sequence in a diploid genome corresponding to a mutation rate of  $2 \times 10^{-8}$  per cell division per site. For the selection models, we assume  $u_b=10^{-5}$  per cell division for driver mutations, on order with that previously suggested by Bozic *et al.*<sup>50</sup>. We also investigated the impact of a lower selectively advantageous mutation rate ( $u_b=10^{-6}$ ) on the SFS, as this mimics late arising driver mutations (**Supplementary Figure 10**).

We also sought to explore how a naïve model of neutral cancer stem cell (neutral-CSC) driven tumor growth would influence the resultant SFS. Here, each deme comprises two subpopulations – stem cells (SCs) and non-SCs where the SC fraction is  $p(\text{SC})$ . In each cell generation, SCs divide symmetrically generating two SCs with probability  $\alpha$  and asymmetrically generating one SC and one non-SC with probability  $\beta$  (where  $\alpha+\beta=1$  and thus the probability of symmetric SC differentiation is 0). Non-SCs can only divide with probability  $\gamma$  or die with probability  $\delta$  (where  $\gamma+\delta=1$ ). We exploit a set of parameter values: namely  $\alpha=0.15$ ,  $\beta=0.85$ ,  $\gamma=0.565$  and  $\delta=0.435$  to ensure the maximum deme size is  $\sim 10,000$  cells and the SC fraction  $p(\text{SC}) \approx 1\text{-}2\%$ , consistent with estimates in solid tumors<sup>51</sup>. While it is of potential interest to consider a CSC model in the context of selection, this is complicated by the need for additional parameters with little experimental support, and hence we do not investigate this here.

During virtual tumor growth, each mutation was assigned a unique index and is recorded with respect to its genealogy and host cells during the simulation, enabling analysis of its frequency in a subpopulation or the whole tumor at different stages of growth. Once the tumor reached a final size of  $\sim 10^9$  cells, approximately the size when it is detectable and routinely resected, we virtually sampled: 1, 2, 4, or 8 regions composed of  $\sim 10^6$  cells from an individual virtual tumor (200 tumors under each of the 7 evolutionary modes, totalling 1400 virtual tumors). The VAF of all SSNVs in the sampled bulk subpopulation were considered the true value, whereas observed VAF values were obtained via a statistical model that mimics the random sampling of alleles during sequencing. In particular, we applied a Binomial distribution ( $n, f$ ) to generate the observed VAF of each site given its true frequency  $f$  and number of covered reads  $n$ . The number of covered reads in each site is assumed to follow a negative-binomial distribution. Here, we assume depth=80 representing 80x sequencing depth on average with a variation in parameter size of 2. A mutation is called when the number of variant reads is  $\geq 3$ , thereby applying the same criteria as for the actual tumors. For each virtual tumor, 100 clonal SSNVs were assigned to represent public mutations, where VAF values were simulated using the statistical model described above with mean VAF of 0.5.

### Identification of subclonal SSNVs in MRS

A SSNS  $m$  is defined as subclonal if all of the three following criterion are met,

- 1) A total probability  $P_m = \prod_{i=1}^k P_{mi} (X_{mi} \leq S_{mi}, N_{mi}, f, \text{pub}_{mi}) < 0.05$ , where  $P_{mi}$  is a binomial probability for region  $i$  of observing less than or equal to  $S_{mi}$  reads carrying mutant allele out of total reads  $N_{mi}$ , provided a lower bound of expected allele frequency if  $m$  is public, given that the tumor purity for region  $i$  equals to

$pu_i$ , the total, minor copy numbers and the cellular prevalence of the SCNA where  $m$  resides equal to  $nt_{mi}$ ,  $nb_{mi}$ , and  $pa_{mi}$  within the tumor content,

$$f.pub_{mi} = \begin{cases} pu_i \times \frac{nb_{mi}}{nc_{mi}} & \text{if } nb_{mi} \geq 1, nc_{mi} \geq 2 \\ (pu_i \times (nt_{mi} - nb_{mi})) / nc_{mi} & \text{otherwise} \end{cases}$$

where  $nc_{mi} = nt_{mi} \times pa_{mi} \times pu_i + 2 \times (1 - pa_{mi} \times pu_i)$ . For sites devoid of SCNAs,  $nt_{mi} = 2, nb_{mi} = 1$  and  $pa_{mi} = 0$ .

- 2) At least one region  $i$  with  $CCF_{mi} \pm 95\% CI_{mi} < 1$
- 3) At least one region  $i$  with adjusted VAF  $VAF_{ami} < 0.25$ . Here 0.25 was chosen because of its good performance in defining subclonality based on simulated virtual tumors (**Supplementary Figure 36**).

A SSNV that does not meet one of the above criterion is considered public. SSNVs with varying patterns of loss of heterozygosity (LOH) amongst regions were not included for pairwise SFS comparisons. The pooled cumulative SFS was computed when multiple samples were available. Here we employ an  $f\_max$  of 0.25 as the upper value for subclonal mutations, whereas  $f\_min$  depends on the total sequencing depth (and hence number of regions sequenced) and is chosen conservatively, while maximizing the inclusion of high confidence low VAF SSNVs.

### ITH metrics

For pairwise comparisons between regions, subclonal (private) SSNVs were assigned as being either private-shared or region-specific. Private-shared SSNVs are present in both regions, whereas region-specific SSNVs are unique to one region where we reject a null model of the same VAF in the other region (given the sequencing depth) with a 5% significance level. For each pairwise SFS histogram, the bin width was optimized for visualization purposes based on the number of SSNVs<sup>52</sup>. Metrics capturing between-region ITH were computed for  $k$  regions and  $r = \binom{k}{2}$  pairwise comparisons as follows:

- 1)  $fHsub = \frac{1}{k} \times \sum_{i=1}^k \frac{SM_i^{high}}{SM_i^{all}}$ , where  $SM_i^{high}$ ,  $SM_i^{all}$  are the number of high frequency subclonal SSNVs (adjusted VAF>0.2, hereafter referred to as VAF) and the number of all subclonal SSNVs with VAF>0.08 for region  $i$ . The cutoff was set to 0.2 since above this value  $fHsub$  tends to plateau in its sensitivity to distinguish the neutral and selection models (**Supplementary Figure 36**). A lower cutoff of 0.08 was chosen empirically to satisfy the tradeoff between the number of subclonal SSNVs and variant calling errors.

- 2)  $fHrs = \frac{1}{r} \times \sum_{j=1}^r \left( \frac{RSM_{ja}^{high}}{2 \times RSM_{ja}^{all}} + \frac{RSM_{jb}^{high}}{2 \times RSM_{jb}^{all}} \right)$ , where  $RSM_{ja}^{high}$ ,  $RSM_{ja}^{all}$  represent the number of high-frequency (VAF>0.2) region-specific SSNVs and the number of all region-specific SSNVs with VAF>0.08 for region  $a$ , in a pairwise comparison  $j$  between regions  $a$  and  $b$ .

- 3)  $FST = \frac{1}{r} \times \sum_{j=1}^r FST_j^{hudson}$ , and  $FST_j^{hudson} = \frac{\sum_{m=1}^{m_t} (f_a^m - f_b^m)^2 - \frac{f_a^m \times (1-f_a^m)}{d_a^m - 1} - \frac{f_b^m \times (1-f_b^m)}{d_b^m - 1}}{\sum_{m=1}^{m_t} f_a^m \times (1-f_b^m) + f_b^m \times (1-f_a^m)}$ , where  $f_a^m$  is the VAF for SSNV  $m$  and  $d_a^m$  is the sequencing depth for SSNV  $m$  in region  $a$ . The genetic variance components (nominator and denominator) are

averaged separately to obtain a ratio combining the Hudson FST estimates across all  $m_t$  SSNVs<sup>53</sup>.

- 4)  $KSD = \frac{1}{r} \times \sum_{j=1}^r KSD_j$ , and  $KSD_j = \max|F_a - F_b|$ , where  $F_a$  is the cumulative SFS of region  $a$ , in a pairwise comparison  $j$  between regions  $a$  and  $b$ .
- 5)  $rAUC = \frac{AUC_{merged}}{AUC_{theor}}$ , corresponding to the ratio of the area under the pooled cumulative SFS to the area under a theoretical cumulative SFS assuming neutral exponential growth of a well-mixed population<sup>21,26</sup>. For MRS, the pooled VAF is the total number of alternative alleles divided by total read depth. As this represents the alternative allele frequency pooled across tumor regions, it should capture overall tumor dynamics, but not between region diversity and complements other ITH metrics.

To evaluate the power (at a significance level of 0.10) or sensitivity of ITH metrics to distinguish a specific alternate model from the neutral model in the simulated data given varying numbers of samples ( $n=1, 2, 4$  and  $8$ ) or variable sequencing depths of a single sample ( $80-640\times$ ), we employed  $rAUC$  as it is applicable to single sample data, as well as  $fHsub$ , one of the MRS specific statistics. The power was computed empirically as the percentage of virtual tumors under an alternative model for which the statistic ( $rAUC$  or  $fHsub$ ) was greater than 95% or less than 5% of the corresponding statistic in the neutral model (taking the larger percentage).

### Evolutionary mode classifier

A radial basis function (RBF) kernel SVM was built based on 1,400 simulated tumors derived from seven growth models (200 for each of neutral, neutral-CSC,  $s=0.01, 0.02, 0.03, 0.05$  and  $0.1$ ). We grouped virtual tumors simulated under the neutral, neutral-CSC and  $s=0.01$  models as "effectively-neutral" and those simulated under higher selection coefficients ( $s \geq 0.02$ ) as "selection" based on the distribution of the five statistics (**Figure 6a**). The five ITH metrics derived from the SFS were Z-score centered and scaled to have mean 0 and SD equal to 1. The SVM was trained using 10 fold cross validation with the R package *caret*<sup>54</sup>. Two rounds of training were performed to optimize the two parameters for RBF ( $C$ : the "cost" of the radial kernel and  $\sigma$ : the smoothing parameter). In the first round, tuning parameters were arbitrarily selected and the default settings were used for the remainder. The training function was employed to calculate estimates for the parameters. In a second round, sensitivity analysis was performed to refine the parameter choice. To evaluate the relative importance of different combinations of the five ITH metrics for classification, SVMs were run 20 times for each of 26 possible combinations of five statistics with the same seed used for random splitting, where 4/5 virtual tumors were used for training and 1/5 for testing, and the resulting ROC AUCs were compared (**Supplementary Figure 27**). A SVM was also built using the two major independent components (IC) obtained from independent component analysis (ICA) of the five ITH metrics where the decision boundaries are shown on the ICA scatter plots. ICA was performed on features derived from the virtual tumors and patient tumors for  $n=2$  (**Supplementary Figure 28**),  $n=4$  (**Figure 6b**) and  $n=8$  (**Supplementary Figure 29**) virtual tumor regions. The performance of the SVM to distinguish each alternative model from the neutral model was evaluated by comparing 100 virtual tumors for training and 100 virtual tumors for testing (**Supplementary Figure 26**).

### Functionality assessment of private and public SSNVs

The ratio of private SSNVs at more functional (MF) relative to less functional (LF) sites was determined as previously described<sup>40</sup> in order to evaluate the correlation between dMF/dLF and various ITH metrics derived from the SFS. SSNVs were considered MF if classified by Polyphen-2 as "damaging" or "probably damaging" and LF if classified as "benign". The dMF/dLF ratio was calculated by normalizing MF/LF for private SSNVs in each tumor to a background MF/LF ratio based on random substitutions in the mutated genes. We also determined the fold enrichment for driver genes (defined based on IntOGen v.2016.5) amongst non-silent public SSNVs and the correlation with various ITH metrics.

### Code availability

Code for the simulation studies and the Variant Assurance Pipeline are available at: <https://github.com/cancersysbio/VirtualTumorEvolution>  
<https://github.com/cancersysbio/VAP>

### Data availability

The single gland WES data and xenograft WES data are available at EMBL-EBI ArrayExpress under accession number E-MTAB-5547. Data from previously published studies are available at: dbGAP: phs000178/GRU; European Genotype Phenotype Archive (EGA): EGAD00001001394, EGAD00001000714, EGAD00001000900, EGAD00001000984, EGAD00001001113; EMBL-EBI European Nucleotide Archive (ENA): PRJEB12737

## References

1. Nordling, C.O. A new theory on cancer-inducing mechanism. *Br J Cancer* **7**, 68-72 (1953).
2. Armitage, P. & Doll, R. The age distribution of cancer and a multi-stage theory of carcinogenesis. *Br J Cancer* **8**, 1-12 (1954).
3. Nowell, P.C. The clonal evolution of tumor cell populations. *Science* **194**, 23-8 (1976).
4. Cairns, J. Mutation selection and the natural history of cancer. *Nature* **255**, 197-200 (1975).
5. Fearon, E.R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759-67 (1990).
6. Tsao, J.L. *et al.* Genetic reconstruction of individual colorectal tumor histories. *Proc Natl Acad Sci U S A* **97**, 1236-41 (2000).
7. Tomasetti, C., Vogelstein, B. & Parmigiani, G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. *Proc Natl Acad Sci U S A* **110**, 1999-2004 (2013).
8. Hu, Z., Sun, R. & Curtis, C. A population genetics perspective on the determinants of intra-tumor heterogeneity. *Biochim Biophys Acta* (2017).
9. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-58 (2013).
10. Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007 (2012).
11. Fischer, A., Vazquez-Garcia, I., Illingworth, C.J. & Mustonen, V. High-definition reconstruction of clonal composition in cancer. *Cell Rep* **7**, 1740-52 (2014).
12. Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nat Methods* **11**, 396-8 (2014).

13. Miller, C.A. *et al.* SciClone: inferring clonal architecture and tracking the spatial and temporal patterns of tumor evolution. *PLoS Comput Biol* **10**, e1003665 (2014).
14. Deshwar, A.G. *et al.* PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol* **16**, 35 (2015).
15. Sottoriva, A. *et al.* A Big Bang model of human colorectal tumor growth. *Nature Genetics* **47**, 209-16 (2015).
16. Uchi, R. *et al.* Integrated Multiregional Analysis Proposing a New Model of Colorectal Cancer Evolution. *Plos Genetics* **12**(2016).
17. Sievers, C.K. *et al.* Subclonal diversity arises early even in small colorectal tumours and contributes to differential growth fates. *Gut* (2016).
18. Bozic, I., Gerold, J.M. & Nowak, M.A. Quantifying Clonal and Subclonal Passenger Mutations in Cancer Evolution. *PLoS Comput Biol* **12**, e1004731 (2016).
19. Suzuki, Y. *et al.* Multiregion ultra-deep sequencing reveals early intermixing and variable levels of intratumoral heterogeneity in colorectal cancer. *Mol Oncol* **11**, 124-139 (2017).
20. Ling, S. *et al.* Extremely high genetic diversity in a single tumor points to prevalence of non-Darwinian cell evolution. *Proc Natl Acad Sci U S A* **112**, E6496-505 (2015).
21. Williams, M.J., Werner, B., Barnes, C.P., Graham, T.A. & Sottoriva, A. Identification of neutral tumor evolution across cancer types. *Nat Genet* (2016).
22. Bustamante, C.D., Wakeley, J., Sawyer, S. & Hartl, D.L. Directional selection and the site-frequency spectrum. *Genetics* **159**, 1779-88 (2001).
23. Ray, N., Currat, M. & Excoffier, L. Intra-deme molecular diversity in spatially expanding populations. *Mol Biol Evol* **20**, 76-86 (2003).
24. Siegmund, K. & Shibata, D. At least two well-spaced samples are needed to genotype a solid tumor. *BMC Cancer* **16**, 250 (2016).
25. Holsinger, K.E. & Weir, B.S. Genetics in geographically structured populations: defining, estimating and interpreting  $F_{ST}$ . *Nat Rev Genet* **10**, 639-50 (2009).
26. Durrett, R. Population Genetics of Neutral Mutations in Exponentially Growing Cancer Cell Populations. *Ann Appl Probab* **23**, 230-250 (2013).
27. Kimura, M. Model of effectively neutral mutations in which selective constraint is incorporated. *Proc Natl Acad Sci U S A* **76**, 3440-4 (1979).
28. Ohta, T. & Gillepsie, J.H. Development of neutral and nearly neutral theories. *Theor. Popul Biol* **49**, 128-142 (1996).
29. Rowan, A. *et al.* Refining molecular analysis in the pathways of colorectal carcinogenesis. *Clin Gastroenterol Hepatol* **3**, 1115-23 (2005).
30. Qiao, Y. *et al.* SubcloneSeeker: a computational framework for reconstructing tumor clone structure for cancer variant interpretation and prioritization. *Genome Biol* **15**, 443 (2014).
31. Li, B. & Li, J.Z. A general framework for analyzing tumor subclonality using SNP array and DNA sequencing data. *Genome Biol* **15**, 473 (2014).
32. Popic, V. *et al.* Fast and scalable inference of multi-sample cancer lineages. *Genome Biol* **16**, 91 (2015).
33. Ross-Innes, C.S. *et al.* Whole-genome sequencing provides new insights into the clonal architecture of Barrett's esophagus and esophageal adenocarcinoma. *Nat Genet* **47**, 1038-46 (2015).
34. Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256-9 (2014).



35. de Bruin, E.C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-6 (2014).
36. Johnson, B.E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189-93 (2014).
37. Kim, H. *et al.* Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res* **25**, 316-27 (2015).
38. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585-95 (1989).
39. Grossman, S.R. *et al.* A composite of multiple signals distinguishes causal variants in regions of positive selection. *Science* **327**, 883-6 (2010).
40. Ostrow, S.L., Barshir, R., DeGregori, J., Yeger-Lotem, E. & Hershberg, R. Cancer evolution is associated with pervasive positive selection on globally expressed genes. *PLoS Genet* **10**, e1004239 (2014).
41. Messer, P.W. & Petrov, D.A. Population genomics of rapid adaptation by soft selective sweeps. *Trends Ecol Evol* **28**, 659-69 (2013).
42. Lloyd, M.C. *et al.* Darwinian Dynamics of Intratumoral Heterogeneity: Not Solely Random Mutations but Also Variable Environmental Selection Forces. *Cancer Res* **76**, 3136-44 (2016).
43. McFarland, C.D., Korolev, K.S., Kryukov, G.V., Sunyaev, S.R. & Mirny, L.A. Impact of deleterious passenger mutations on cancer progression. *Proc Natl Acad Sci U S A* **110**, 2910-5 (2013).
44. Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54-8 (2014).
45. Cleary, A.S., Leonard, T.L., Gestl, S.A. & Gunther, E.J. Tumour cell heterogeneity maintained by cooperating subclones in Wnt-driven mammary cancers. *Nature* **508**, 113-7 (2014).
46. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol* **31**, 213-9 (2013).
47. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res* **24**, 1881-93 (2014).
48. Waclaw, B. *et al.* A spatial model predicts that dispersal and cell turnover limit intratumour heterogeneity. *Nature* **525**, 261-4 (2015).
49. Diaz, L.A., Jr. *et al.* The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers. *Nature* **486**, 537-40 (2012).
50. Bozic, I. *et al.* Accumulation of driver and passenger mutations during tumor progression. *Proc Natl Acad Sci U S A* **107**, 18545-50 (2010).
51. Visvader, J.E. & Lindeman, G.J. Cancer stem cells in solid tumours: accumulating evidence and unresolved questions. *Nat Rev Cancer* **8**, 755-68 (2008).
52. Wand, M.P. Data-Based Choice of Histogram Bin Width. *The American Statistician* **51**, 59 (1997).
53. Bhatia, G., Patterson, N., Sankararaman, S. & Price, A.L. Estimating and interpreting FST: the impact of rare variants. *Genome research* **23**, 1514-21 (2013).
54. Kuhn, M. Building Predictive Models in R Using the caret Package. *Journal of Statistical Software* **28**, 1-26 (2008).

## **Acknowledgments**

This work was funded by award from the NIH (R01CA182514), Susan G. Komen Foundation (IIR13260750), and the Breast Cancer Research Foundation (BCRF-16-032) to C.C and an award from the NIH (R01CA185016) to D.S. Z.H is supported by an Innovative Genomics Initiative (IGI) Postdoctoral Fellowship. A.S is supported by the Chris Rokos Fellowship. This work was supported in part by NIH P30 CA124435 utilizing the Genetics Bioinformatics Service Center within the Stanford Cancer Institute Shared Resource. The results are in part based upon data generated from the following studies: EGAD00001001394, EGAD00001000714, EGAD00001000900, EGAD00001000984, EGAD00001001113. We thank members of the Curtis lab for helpful discussions.

## **Contributions**

R.S, Z.H, and C.C designed the study. R.S analyzed and visualized the data and performed statistical analyses. Z.H. performed simulation studies. Z.M, D.S generated data. R.S, Z.H, C.C interpreted the data. A.S, T.G contributed to earlier analysis of the COAD dataset. A.H. provided statistical advice. J.M.F performed xenograft experiments. D.S, C.C provided reagents/data. C.C supervised the study and wrote the manuscript with input from R.S and Z.H. All authors read and approved the final manuscript.

## **Competing interests**

The authors declare no competing interests.

## Figures

### Figure 1. Overview of simulation framework and genomic data analysis pipeline.

**(a)** Schematic overview of our agent-based computational framework to simulate 3D tumor growth (after transformation) under various modes of evolution, including neutral evolution (null model) and different levels of positive selection, followed by spatial sampling and multi-region sequencing of the *virtual tumor*. Tumor growth is simulated via the expansion of deme subpopulations within a defined 3D cubic lattice according to explicit rules dictated by spatial constraints, where cells within each deme are well-mixed and grow via a stochastic branching (birth-death) process (Methods and **Supplementary Figure 1**). By simulating the acquisition of random mutations (neutral or beneficial), tracing the genealogy of each cell as the tumor expands and subsequently virtually sampling and sequencing the ‘final’ *virtual tumor* as is done experimentally after resection or biopsy, it is possible to evaluate differences in the site frequency spectrum (SFS) under different modes of selection and sampling strategies. Five Intra-tumor heterogeneity (ITH) metrics derived from the SFS were employed to distinguish between different evolutionary modes. **(b)** A unified sequencing analysis pipeline based on SSNV calling, copy number estimation, as well as stringent quality control was employed to obtain variant allele frequency (VAF) estimates adjusted for purity and local copy number for seven multi-region sequencing (MRS) datasets derived from patient samples across diverse tissue types. The ITH metrics were similarly computed in patient tumor samples and compared to those observed in virtual tumors under different evolutionary modes.

### Figure 2. Characteristics of virtual tumors simulated under different modes of evolution.

**(a)** A 2D visualization of a clone map in *virtual tumors* simulated under different modes of evolution, including the null neutral model (selection coefficient,  $s=0$ ), a neutral model with cancer stem cell driven growth (neutral-CSC), and varying levels of selection ( $s=0.01, 0.05$  and  $0.1$ ). Colors correspond to distinct clones with high VAF ( $> 0.4$ ) in each deme subpopulation. **(b)** Representative pairwise SFS histograms derived from two spatially separated regions (labeled A and B) within the same tumor are shown for tumors simulated under different evolutionary modes. SSNVs were classified as Public (gray), Private (Pvt)-shared (green), or Private-region specific (blue) based on their presence in the virtual MRS data (Methods). The total number of SSNVs detected in each region, as well as three ITH metrics are indicated, namely fHsub, FST, KSD. **(c)** The cumulative SFS derived from virtual tumors (100 shown for each mode) was computed based on the pooled VAF for subclonal SSNVs for four regions in the frequency (f) range 0.02–0.25. Curves are Bezier smoothed. The dashed curve corresponds to the average and the black curve to a theoretical cumulative SFS under neutral exponential growth in a well-mixed population. For each mode, the mean ratio of the area under the cumulative SFS from the *virtual tumors* compared to that of the theoretical cumulative SFS (denoted rAUC) based on 100 virtual tumors is indicated as are the 95% bootstrap confidence intervals.

### Figure 3. Colorectal tumors exhibit patterns of between-region genetic divergence consistent with effectively-neutral growth or selection.

**(a)** Pairwise comparison of SFS histograms from each of three bi-sampled colon adenocarcinomas (COADs) representing the major molecular subgroups, including MSI-H (carcinoma W, right), MSS/CIN+ (carcinoma U, middle) and MSS/CIN- (carcinoma M, left). The pairwise histograms illustrate the number of SSNVs detected at a given VAF for the two tumor

regions shown above and below the x-axis. SSNVs were classified as Public (gray), Private (Pvt)-shared (green), or Private-region specific (blue). The total number of SSNVs detected in each region, as well as fHsub, FST, KSD values are indicated. **(b)** Scatterplots comparing SSNVs detected in each tumor region at a given VAF. The color of individual SSNV points corresponds to that in Panel A, and hues reflect the number of SSNVs in a square (0.02 on a side) centered on each SSNV, as depicted in the legend. Nonsilent SSNVs in predicted COAD driver genes are denoted by red circles with known drivers labeled. **(c)** Circos plot illustrating the predicted absolute total CN (Nt) and minor allele CN (Nb) for each tumor sample. Diploid segments are indicated in white for Nt (two copies) and Nb (one copy), while segments with copy number gain and loss are shown in red and blue, respectively, according to the scale bar. Tumor cell purity (Pu) as well as ploidy (Pi) estimates for each region are indicated on the corresponding concentric rings.

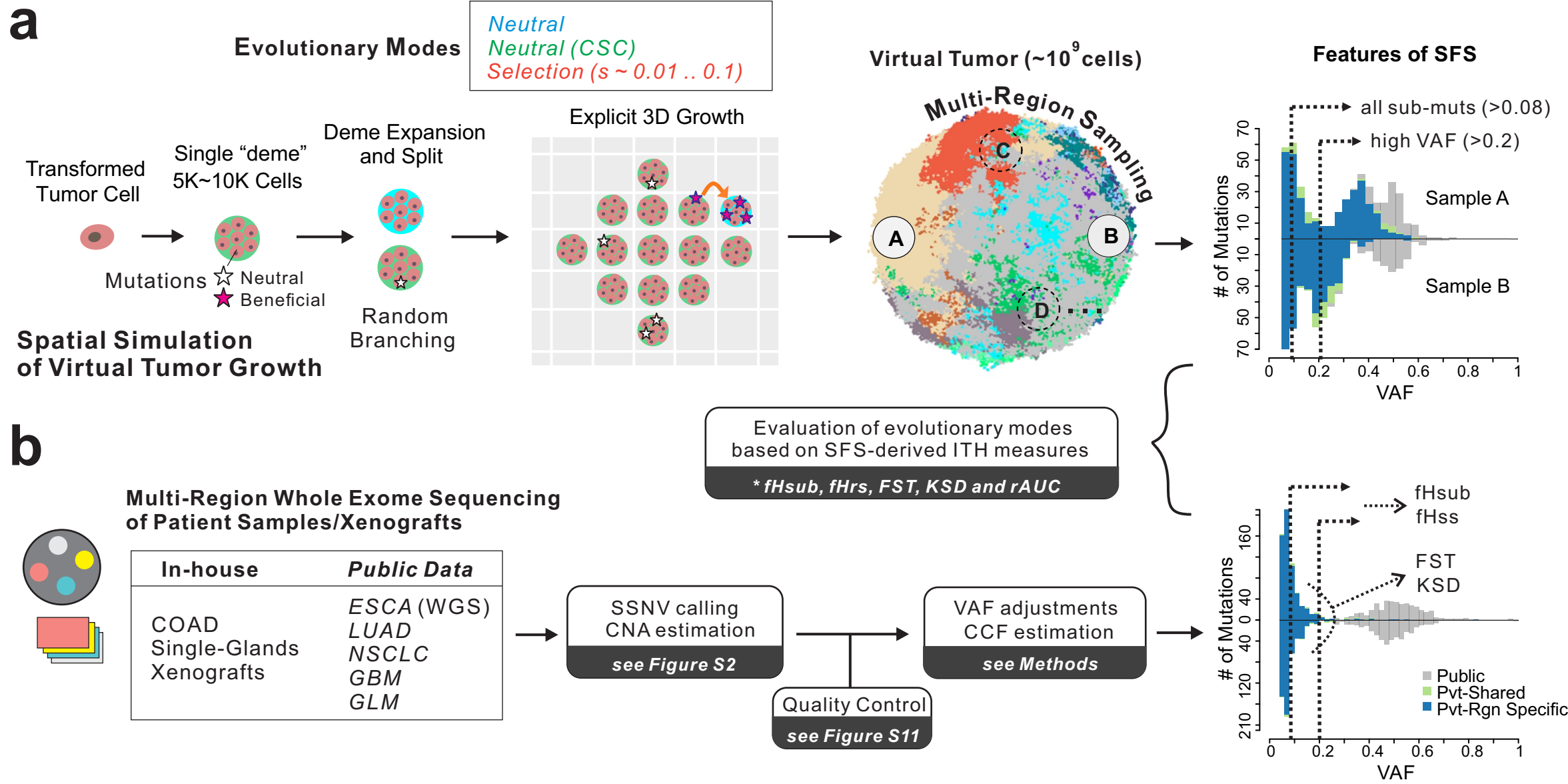
**Figure 4. Single-gland WES reveals spatial constraints amongst subclonal mutations.** **(a)** Pairwise histogram of the SFS and SSNV scatterplots from two regions of COAD-O (OA vs. OB). **(b)** Intersection of SSNVs found in bulk regions and single-glands. In the inset, the VAFs for single-gland vs. bulk sample OA (side-A) specific SSNVs are shown. OA specific SSNVs present in different sets of single-glands collapse to similar VAF values ( $<0.2$ ) in the bulk sample (blue lines connecting the insert), indicating that mutational clusters do not guarantee clonal identity. **(c)** The subclonal cluster of pooled VAF of LUAD-4990 partitioned into multiple clusters for VAF in two separate regions, whereas the clonal VAF cluster (centered at 0.5) persists as a single cluster, consistent with them being present in all cells. Generally, subclonal clusters derived from 'n' regions (for COAD-O, LUAD-4990 and ESCA-8) separate into additional clusters (with more than 5% SSNVs in the original cluster) when 'n+1' regions are employed in analysis (**Supplementary Figures 21-23**). **(d)** Phylogenetic tree based on SSNV presence/absence in single glands and bulk samples constructed using LICHeE. The bulk sample and corresponding single-glands from the same tumor region share a common lineage relationship, reflective of spatial constraints during tumor expansion. SSNVs in known and candidate driver genes are labeled. A truncal APC indel was also detected, but not used for tree construction.

**Figure 5. The SFS reflects differential modes of evolution within and between tumors types.** **(a)** Cumulative SFS based on the merged VAF for tumors derived from four tissue types (colon, esophageal, lung, brain) analyzed using the VAP (Methods). All samples were subject to WES with the exception of the ESCA/BE cases for which WGS was available. Each line corresponds to a Bezier smoothed curve of the cumulative SFS. Thick gray curves correspond to the theoretical cumulative SFS under neutral exponential growth in a well-mixed population, shown for reference. Dashed lines correspond to comparisons of tumor regions sampled at distinct stages of tumor progression in the same patient, e.g., Barrett's esophagus (BE) versus esophageal carcinoma (ESCA), or treatment naïve primary tumor versus post-treatment (Tx) recurrent brain tumors, both of which represent positive controls for selection. **(b)** Pairwise SFS histograms from representative tumors of different tissue origin depict the number of SSNVs detected at a given VAF for two regions, where SSNVs are grouped into Public (gray), Private (Pvt)-Shared (green) and Private-Region specific (blue) mutations (as in Figure 3). Histogram bin widths were optimized based on the number of SSNVs (Methods). **(c)** Two-way density plots of SSNVs present in each region at a given VAF are shown for two tumors exhibiting signals of selection. Non-silent SSNVs in known and candidate driver genes are labeled. The color scale represents the relative

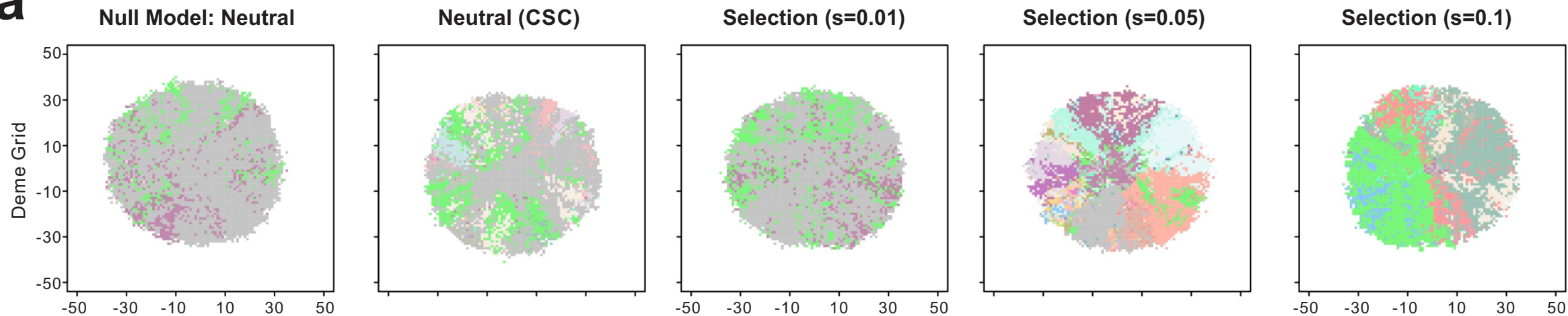
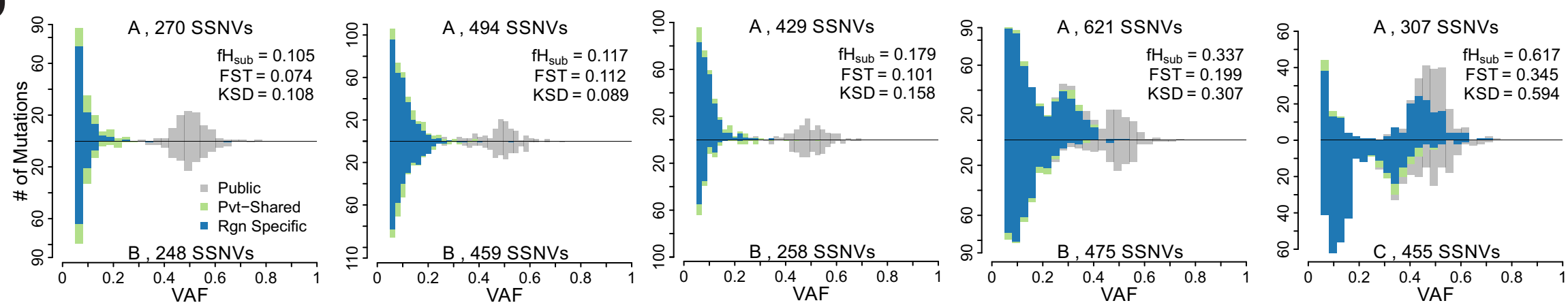
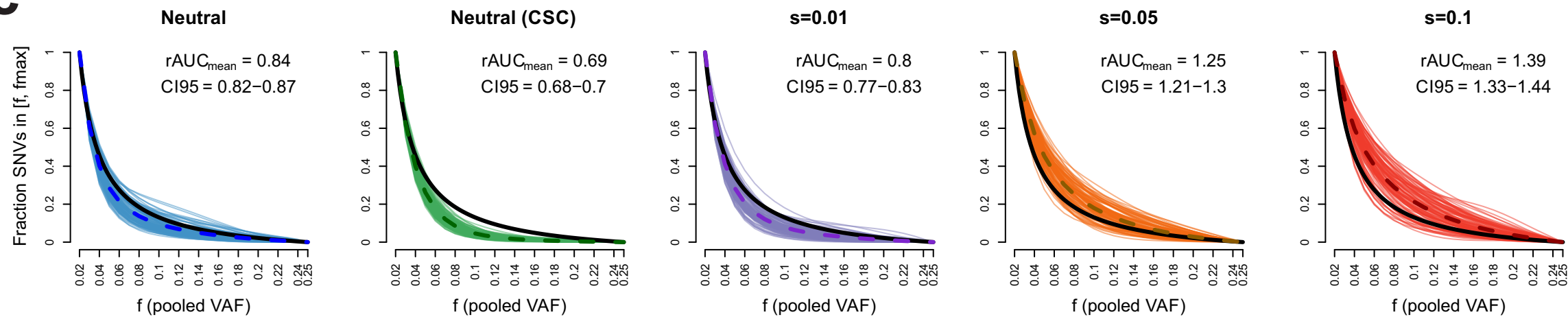
density of mutations.

**Figure 6: Projection of patient samples onto distinct evolutionary modes. (a)** Violin plots for each of five ITH metrics, namely, fHsub, fHrs, Fst, KSD, and rAUC. Colored violin plots show the tumors simulated under different evolutionary modes, whereas the white plots represent real tumor data. Paired pre-treatment primary and post-treatment recurrent brain tumors are denoted by “Tx” and serve as a positive control for selection. **(b)** Independent component analysis (ICA) of simulated and real tumors based on the five ITH metrics. The independent components clearly separate the simulated tumors under *effectively (e) neutral growth* (neutral, neutral-CSC and  $s=0.01$ ) versus *positive selection* ( $s \geq 0.02$ ). The decision boundary for an SVM trained on the two independent components (IC) based on the virtual tumors (*e-neutral* versus *positive selection* models) is indicated by the dashed line. Large transparent colored circles represent values from simulated tumors under different models (200 tumors from each of the seven modes are shown). Small circles corresponding to patient tumors are labeled by their corresponding sample ID, and color-coded according to the nature of the samples. COAD: colorectal adenocarcinoma; CRA: colorectal adenoma; ESCA: esophageal adenocarcinoma; BE: Barrett's esophagus; LUAD: lung adenocarcinoma; NSCLC: non-small-cell lung cancer; GLM: glioma; GBM: glioblastoma; Xeno: COAD cell line xenografts. **(c)** The ratio of private SSNVs at more functional (MF) relative to less functional (LF) sites (dMF/dLF) as defined based on PolyPhen2 was calculated for each of the primary tumors in order to evaluate the correlation with various ITH.

Figure 1

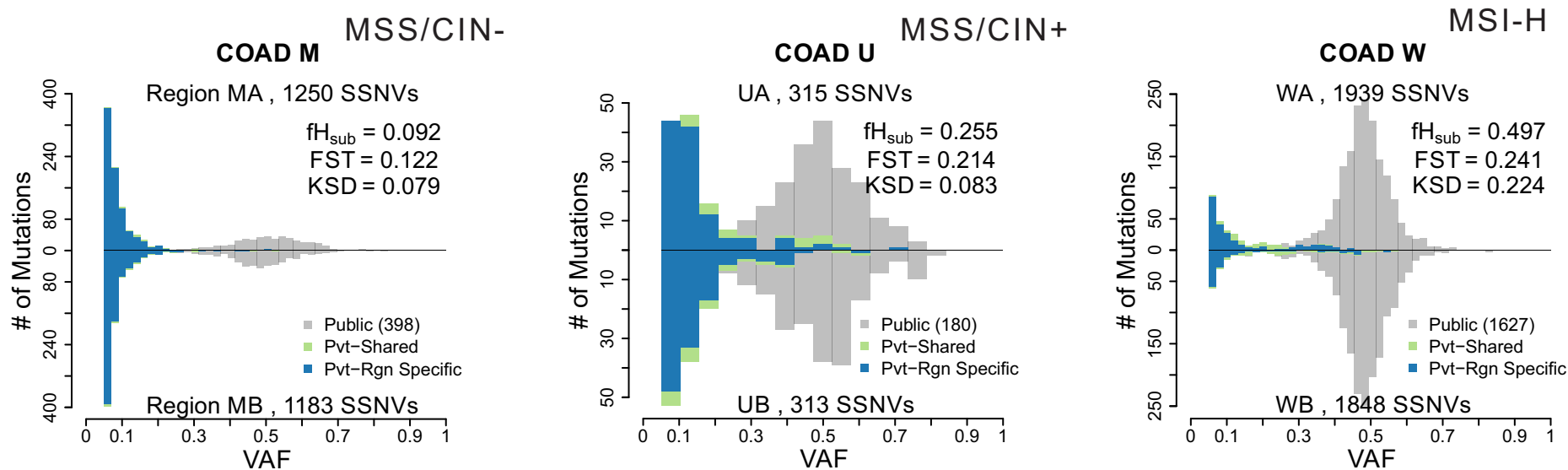


# Figure 2

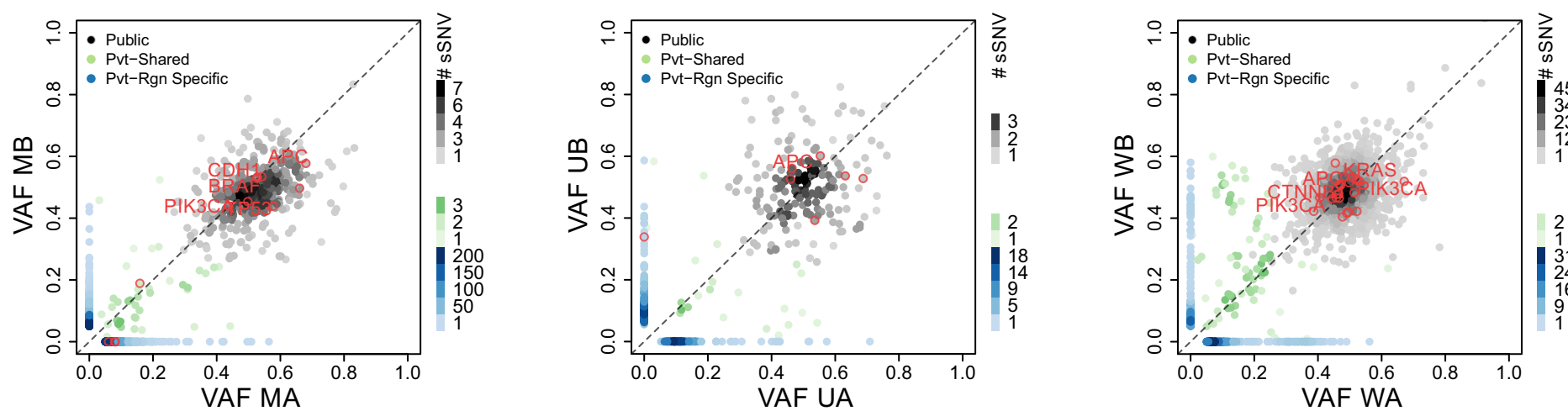
**a****b****c**

# Figure 3

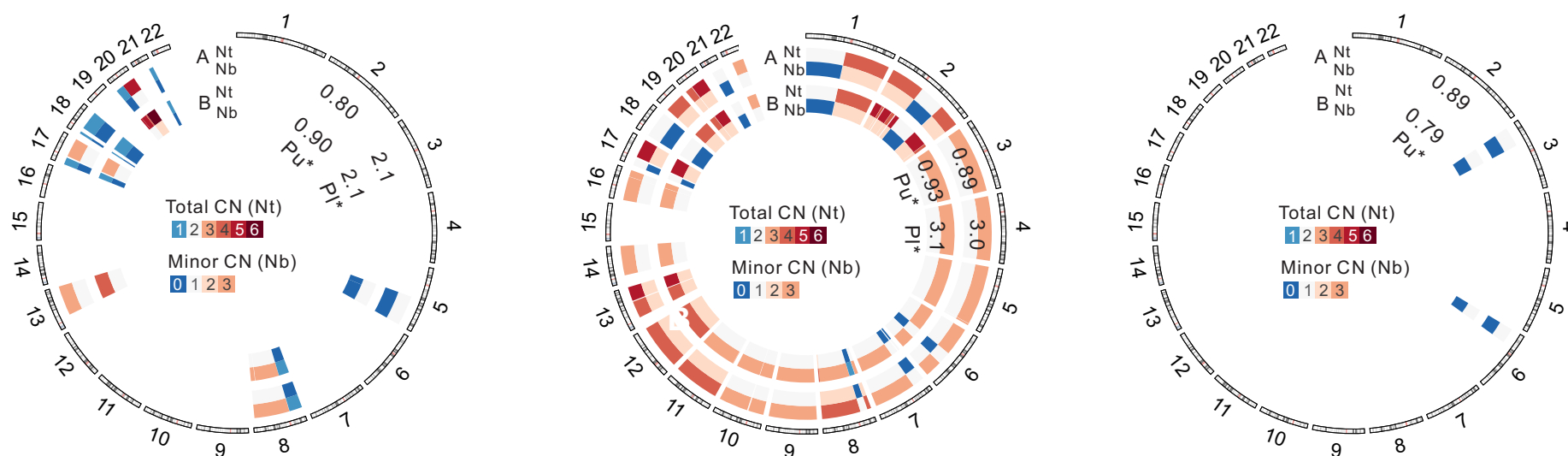
**a**



**b**



**c**





# Figure 4

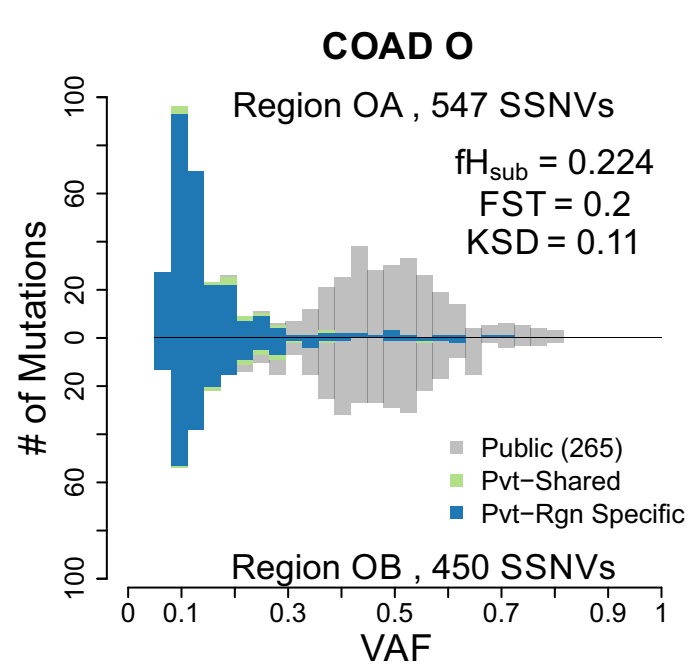
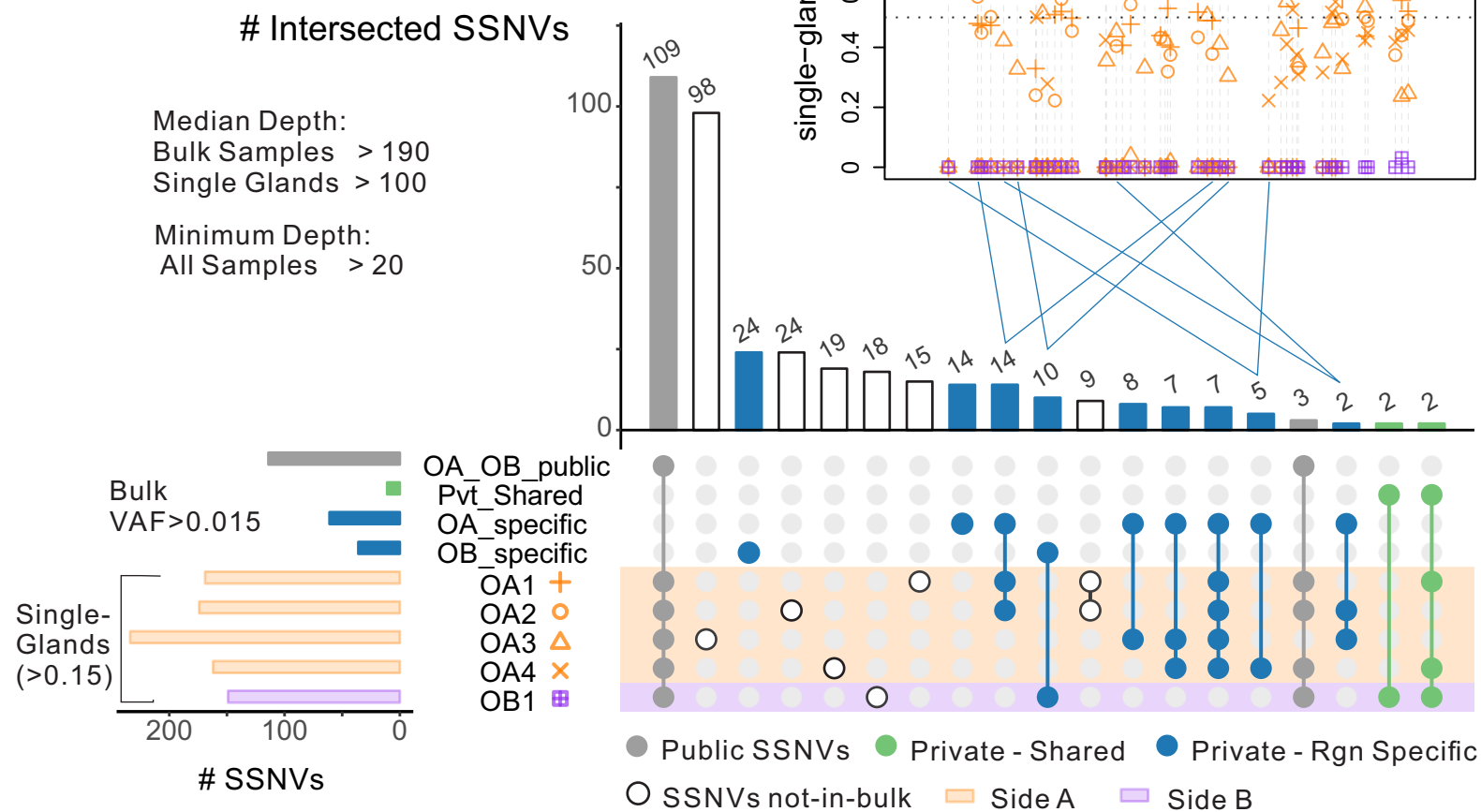
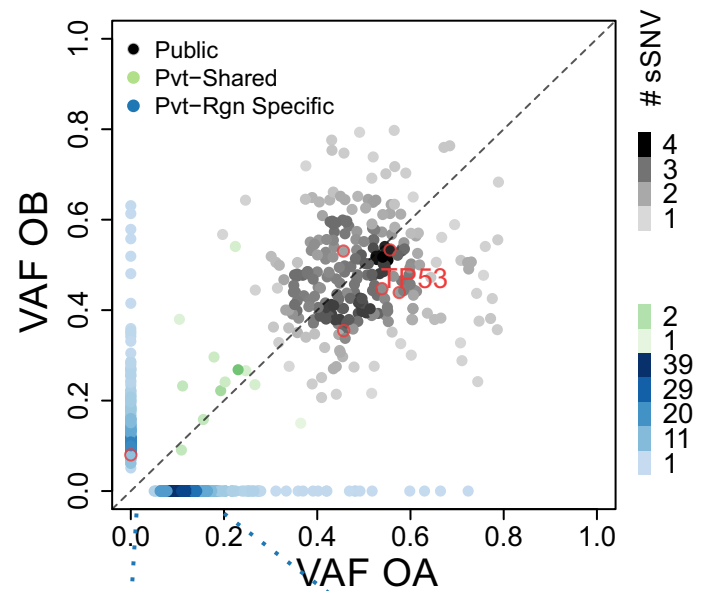
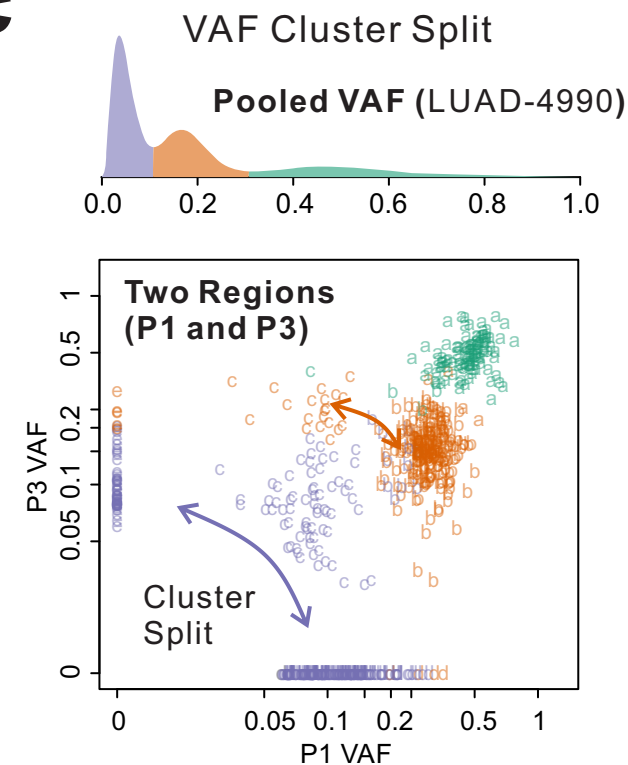
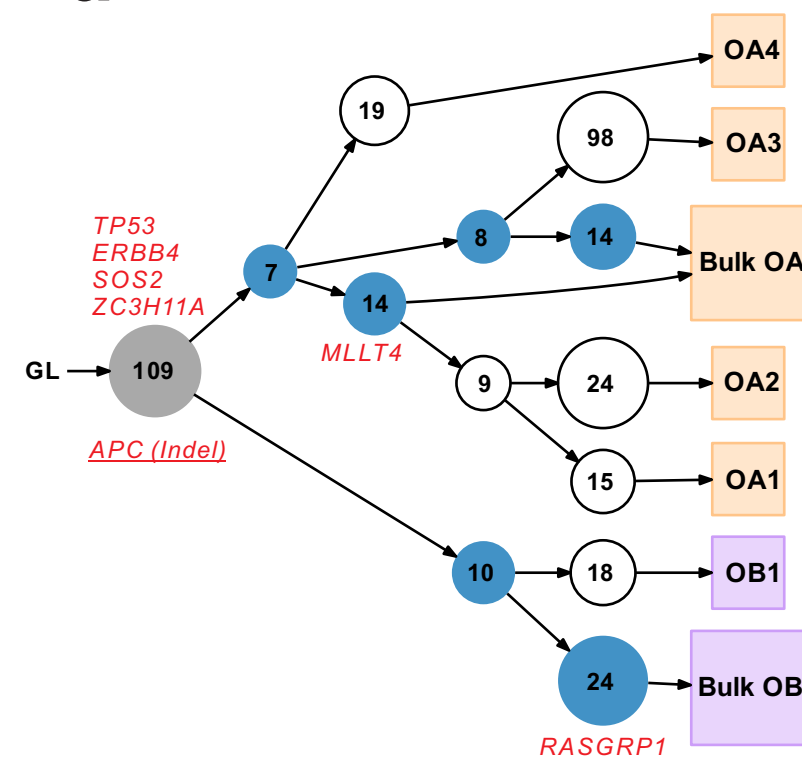
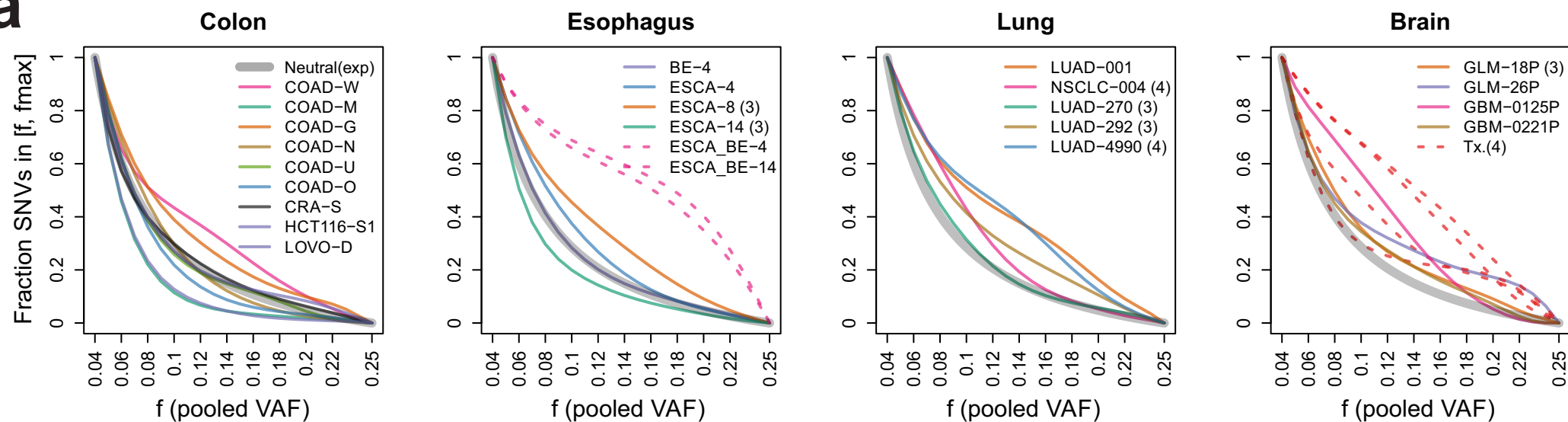
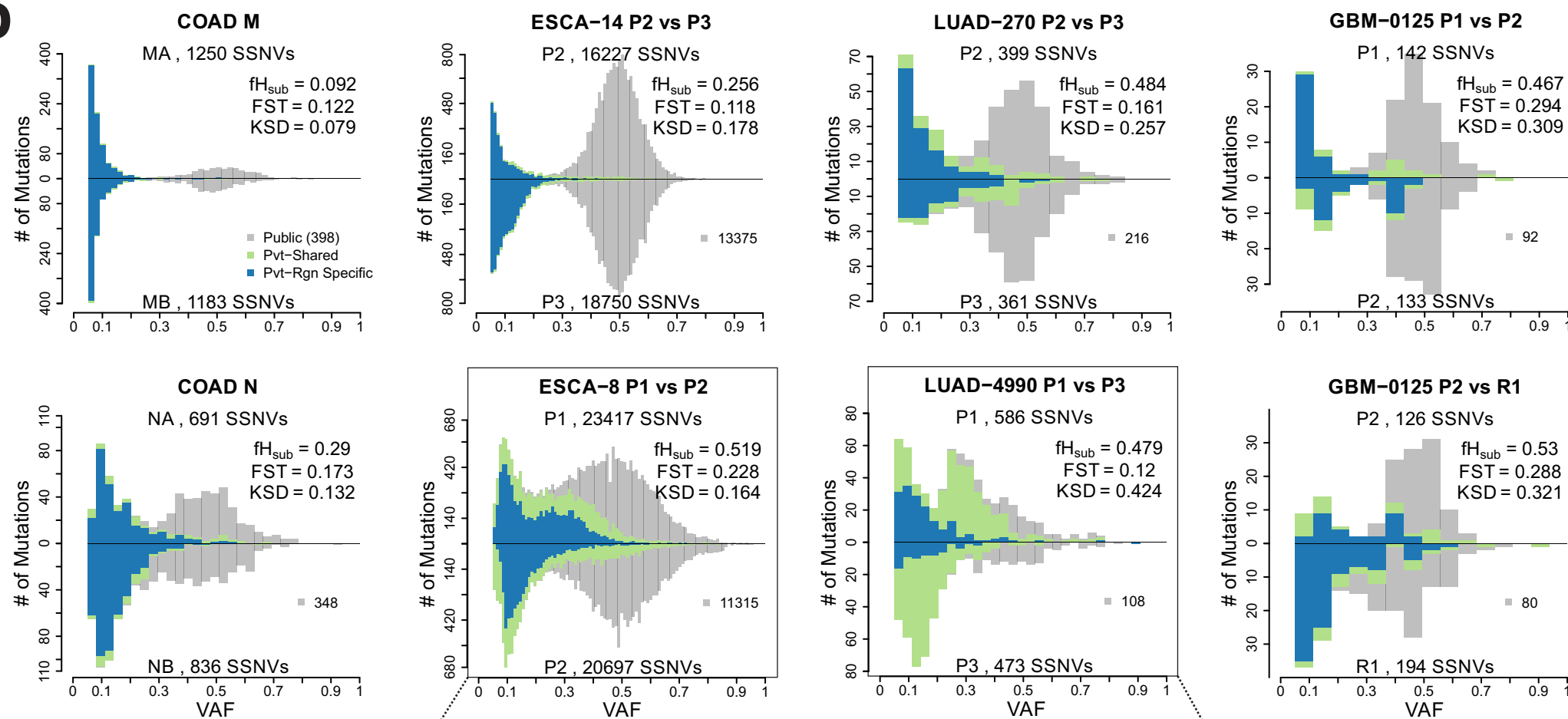
**a****b****COAD O****c****d**

Figure 5

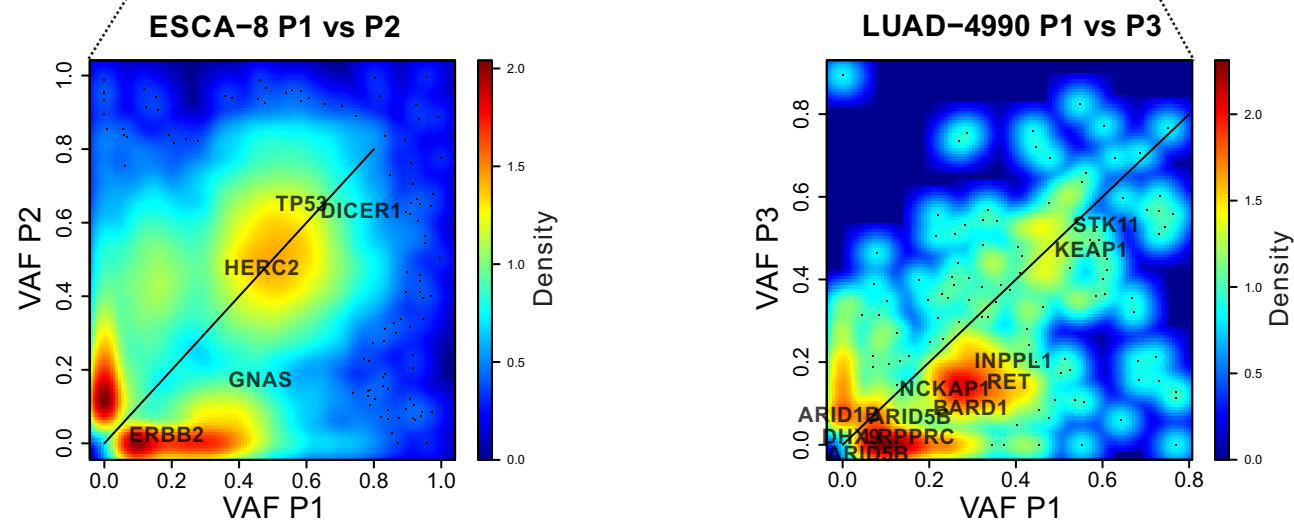
a



b

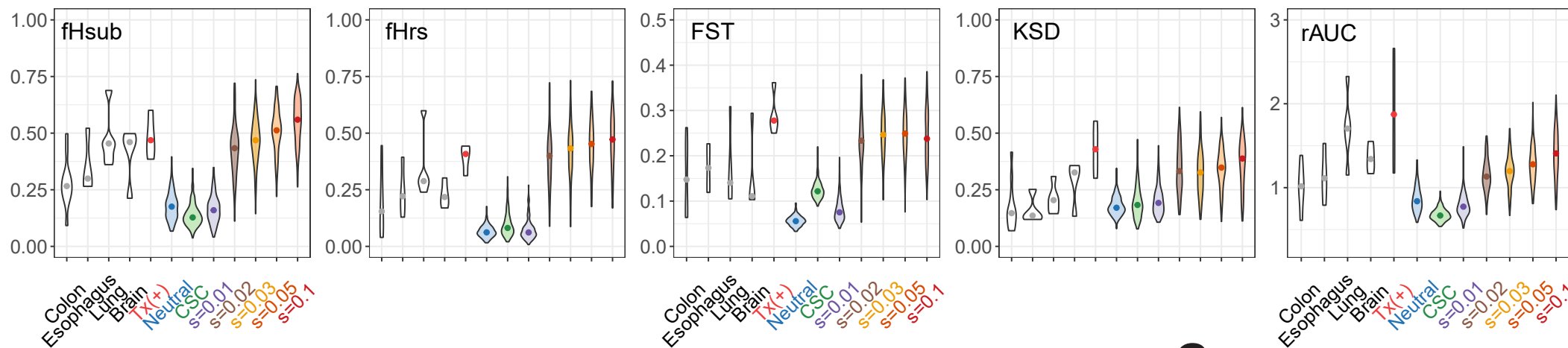


c

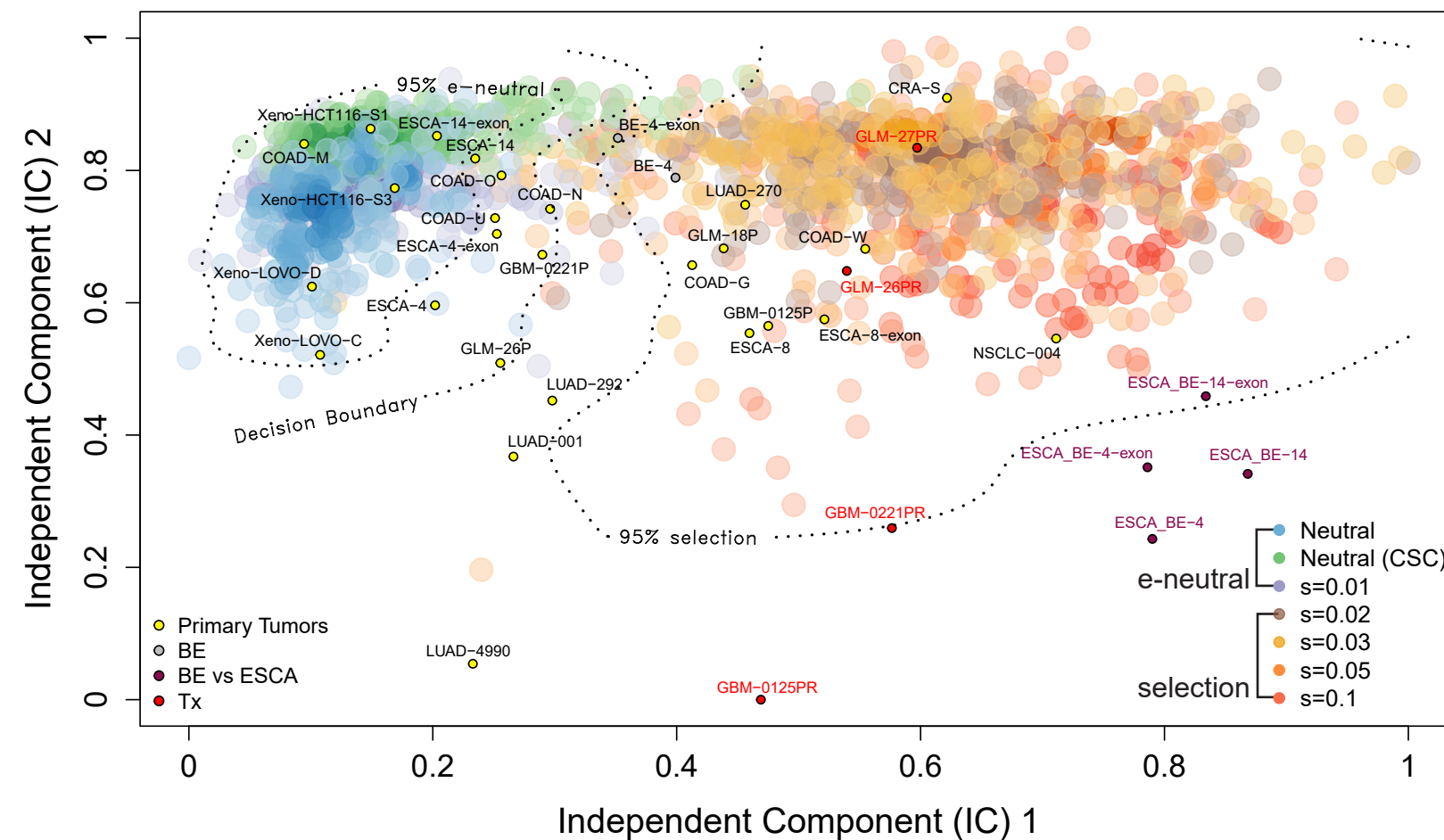


# Figure 6

**a**



**b**



**c**

